



International
Labour
Organization

Evaluation
Office

i-eval THINK Piece, No. 22

Results and reflections from a quality
appraisal of ILO evaluations, 2021

By Universalia Management Group
September 2022



► Results and reflections from a quality appraisal of ILO evaluations, 2021

Think Piece, No. 22

Universal Management Group
ILO Evaluation Office

September 2022

The responsibility for opinions expressed in this document rests solely with the author. The publication does not constitute an endorsement by the International Labour Organization. This document has not been subject to professional editing.

Contents

▶ 1. INTRODUCTION.....	1
▶ 2. FINDINGS	3
▶ 3. CONCLUSIONS	8

► 1. Introduction

1.1 Background

As part of the ILO's Evaluation Office's (EVAL) longstanding quality assurance process¹ the Universalia Management Group Limited (hereinafter, "Universalia" or "the review team") conducted a Quality Appraisal (QA) of 55 independent development cooperation project evaluations submitted in 2021, a sample of ten internal evaluations completed in 2020 and 2021 and the six high-level evaluations completed in 2020 and 2021.

The main purpose of the QA annual report was to provide a cumulative analysis of the evaluations submitted in 2021 and assess trends and comparisons with previous quality appraisals. The Quality Appraisal Summary Report informed the ILO's latest Annual Evaluation Report for 2021-2022, which was released in September 2022.

This Think Piece reports a summary of the findings, conclusions, and recommendations. In addition to the introduction, the report contains two sections:

- Section 2 presents the main findings of the assessment, and
- Section 3 presents the main conclusions of the quality appraisal.

In 2020, the ILO Evaluation Office (EVAL) shifted to a rolling quality assurance process of decentralized project evaluations. Universalia has been appraising these evaluations on a biweekly basis to allow EVAL to identify and act upon emerging quality issues rapidly. Results are disseminated throughout the ILO's evaluation network so that interventions can be made in a targeted manner.

¹ This is one is a series of QAs. A total of ten quality appraisals have been conducted over the years of the independent evaluation reports submitted to EVAL. The results have been described in various Think Pieces, most recently in Robertson and Schroter, ["Leveraging appraisal findings to improve evaluation quality", i-eval Think Piece No.4](#), March 2014; Friedman and Blight, ["External quality appraisal: Implications for evaluation quality and utilization", i-eval Think Piece No.8](#), December 2014; Watts, ["Quality assessments of ILO project evaluations: What are the next steps to better evaluations?", i-eval Think Piece No. 10](#), March 2016 ; Llabres, ["Evaluation quality assessment methodology in the UN system and changes to the ILO's quality appraisal methodology", i-eval Think Piece No. 12](#), December 2017; Bustamante, López and Román, ["ILO evaluation quality: Challenges and potential strategies", i-eval Think Piece No. 13](#), June 2018; Gonzales and Pénicaud, ["Quality assessments of ILO project evaluations: Sustaining recent improvements", i-eval Think Piece No. 17](#), December 2019; Franche, Gonzales and Pénicaud, ["Quality assessments of ILO decentralized evaluations: Key results of the quality appraisal 2019 and way forward for the integration of gender equality and empowerment of women considerations into evaluation", i-eval Think Piece No. 19](#), December 2020.

1.2 Overview and implementation of the quality appraisal tool

The ILO's quality appraisal tool looks at four different dimensions structured in four sections, allowing the reviewers to collect quantitative data on the quality of ILO's evaluation reports. First, the tool captures descriptive data on demographic variables of each evaluation report, such as the region, department, and year. Collected data can be analyzed through the aggregation and identification of trends across these independent variables.

Secondly, the QA tool requires the reviewers to rate the quality of the content of the evaluation reports according to 58 different items (or criteria) grouped across the 10 standard sections that should structure an evaluation report.

Third, the comprehensiveness section of the tool ensures that data is collected on the presence or absence of key components that must be included in the report using a binary scale (absent-present).

Finally, the UN-SWAP assesses four different items, in alignment with the Guidance on Integrating Human Rights and Gender Equality in Evaluation.

The QA report covered 55 independent development cooperation project evaluation reports, prepared for the ILO worldwide, between January and December 2021, a sample of 10 internal evaluations completed in 2020 and 2021, and six high-level evaluations completed over the period 2020-2021. The independent and internal evaluation reports appraised include a mix of final and mid-term evaluations, in English, French and Spanish, covering all regions where the ILO operates and involving several ILO departments and offices.

QA of evaluation reports is a four-step process, involving two reviewers for independent and internal evaluations and three reviewers for HLEs, to ensure inter-rater reliability. Once all reports of a given reporting period were appraised, quantitative ratings and qualitative information justifying the rating were aggregated in an Excel sheet and overall scores were calculated for the quality, comprehensiveness, and UN-SWAP dimensions. Aggregated scores and individual ratings were analyzed using quantitative and qualitative methods. An online survey was also disseminated among 32 Evaluation Managers to which 18 responded.

The review team adopted a concurrent triangulation design as it generated and analyzed the quantitative and qualitative data simultaneously. This method was used to cross-validate findings and to provide plausible explanatory factors for the tendencies that were observed.

The Quality Appraisal process was based on the ILO's evaluation policy, strategy, and guidelines (checklists, etc.) and was aligned with UNEG evaluation norms and standards. In addition, the review team was committed to undertaking a gender-responsive QA by assessing whether gender was adequately mainstreamed across the evaluation reports appraised.

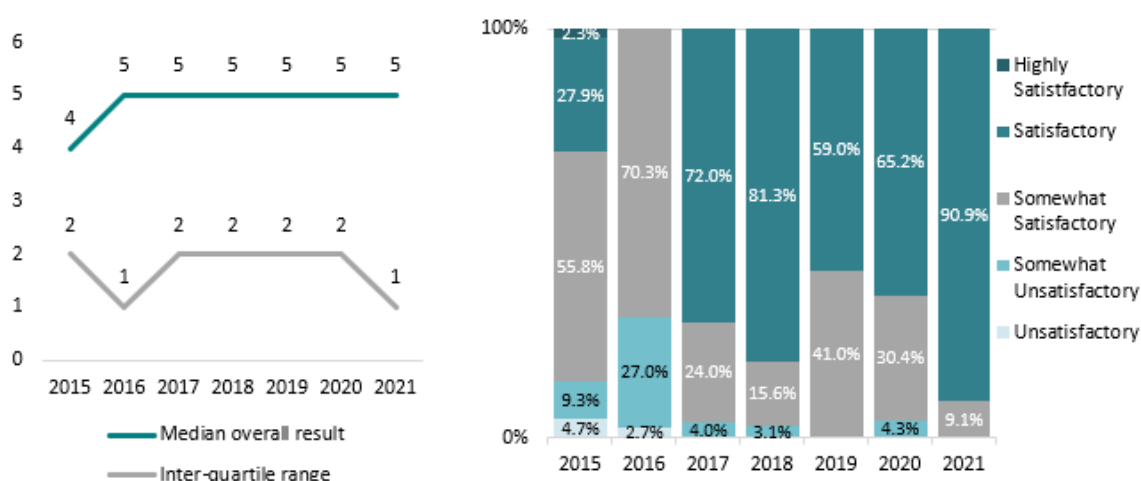
► 2. Findings

2.1 Overall quality score

All appraised independent evaluation reports conducted in 2021 met the required quality standards (“somewhat satisfactory” or “satisfactory” median rating). While 9% of reports obtained a “somewhat satisfactory” rating, 91% obtained a “satisfactory” rating. None of them received an overall assessment of “unsatisfactory”, “highly unsatisfactory” nor “highly satisfactory”. Overall quality of independent evaluation reports has remained satisfactory since 2016.

The dispersion of scores remained low since 2015, suggesting a certain homogeneity in the quality of reports over the years. The inter-quartile range, which measures the dispersion of results around the median, decreased to one in 2021 compared to remaining at two over the period 2017-2020.

Figure 1. Overall ratings and evolution per year²



All independent evaluation reports submitted to EVAL are categorized by department, country, and region. Analysis found that quality is quite consistent across the nine departments and offices represented in the sample, with eight of them obtaining a median score equivalent to a “satisfactory” rating (5 out of 6) for the evaluations that Evaluation Managers oversaw on behalf of EVAL in 2021. The NORMES (n=2) department received a slightly lower median rating corresponding to a “somewhat satisfactory” rating (4 out of 6).

Furthermore, it was found that there is parity in the overall quality of independent evaluation reports, across the regions. Evaluations have a median score considered “satisfactory” (5) in all regions.

² Results for 2015 and 2016 are taken from the QA conducted in 2017 by Artival. Results for 2017 and 2018 are taken from the QA conducted in 2019 by Universalía. Results for 2019 include all the evaluation reports that were reviewed during the previous and the current QA by Universalía.

Of the sample of 10 internal evaluation reports from 2020 and 2021, 70% achieved a “satisfactory” overall rating, while two obtained a “somewhat satisfactory” median rating and one report was an outlier receiving a median score of “highly unsatisfactory”.

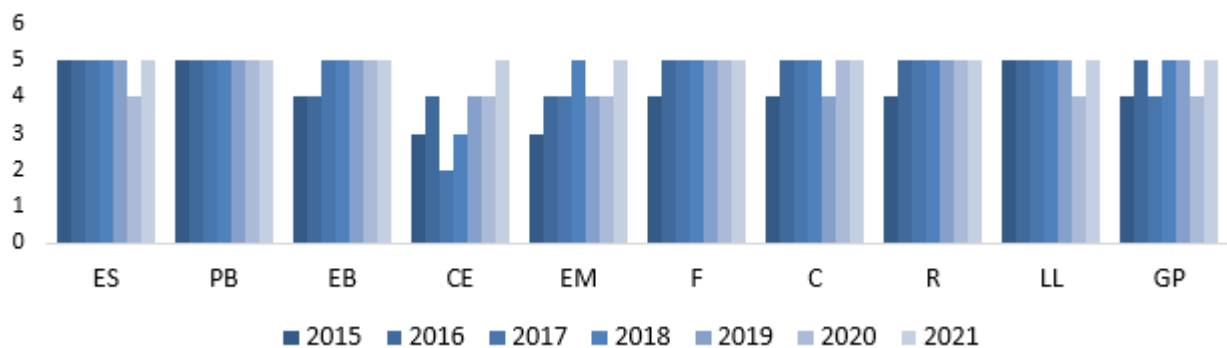
The median score for HLE reports completed in 2020 and 2021 is “satisfactory” for five out of six reports. The inter-quartile range, which is a measure of dispersion around the median, is 1.

2.1 Quality of individual components

This section presents an overview of the ratings obtained per component per year followed by an individual analysis for each individual component of the QA tool.

As illustrated in figure below, all components of independent evaluations obtained “satisfactory” ratings in 2021 (median rating of 5 out of 6). Noticeably, several components received “satisfactory” ratings in 2021 and 2019 but received “somewhat satisfactory” ratings in 2020 (‘Executive Summary’, ‘Lessons Learned’, and ‘Good Practices’). In previous years, the ‘Criteria & Questions’ component was the weakest component. It has improved to the point that it has now received a “satisfactory” rating.

Figure 2. Evolution of median rating per component and year



Executive summary (ES), Project background (PB), Evaluation background (EB), Criteria & questions (CE), Evaluation methodology (EM), Findings (F), Conclusions (C), Recommendations (R), Lessons learned (LL), Good practices (GP).

Internal evaluations achieved “satisfactory” median ratings (5 out of 6) across most components, with the exception of ‘Project Background’ and ‘Evaluation Methods’ obtaining “somewhat satisfactory” (4 out of 6) and ‘Findings’ obtaining a median rating of 4.5.

2.2. Comprehensiveness of evaluation reports

The QA tool includes a section on the comprehensiveness of evaluation reports. The purpose of this section is to assess the overall structure of independent evaluations and to report any missing elements. The reviewers utilize a binary scale to quantify the presence or absence of specific components expected to be found in an evaluation report.

The independent evaluation reports that were appraised generally included EVAL's mandatory components of an evaluation report. As indicated in the table below, the average rating increased in 2021 to 95.1%, the highest average rating in the past seven years. While only 62% of independent evaluations included a 'list of tables, figures, and charts', this is still an improvement on previous years. 'Annexes' are still incomplete for 29% of independent evaluations.

QA results on comprehensiveness of evaluation reports (2021)

COMPONENT	2015	2016	2017	2018	2019	2020	2021
Title page using EVAL's template	67.4%	73.0%	92.0%	90.6%	94.9%	100.0%	100.0%
Table of Contents	97.7%	94.6%	96.0%	100.0%	100.0%	100.0%	100.0%
List of tables, figures and charts	32.6%	37.8%	36.0%	43.8%	53.8%	39.1%	61.8%
List of annexes	79.1%	94.6%	92.0%	87.5%	94.9%	97.8%	96.4%
List of acronyms or abbreviations	100.0%	91.9%	100.0%	93.8%	100.0%	97.8%	100.0%
Executive Summary	100.0%	97.3%	100.0%	96.9%	97.4%	95.7%	98.2%
Project Background	100.0%	97.3%	100.0%	100.0%	100.0%	100.0%	100.0%
Evaluation Background	100.0%	100.0%	100.0%	100.0%	100.0%	97.8%	100.0%
Criteria and Questions	72.1%	83.8%	72.0%	75.0%	82.1%	93.5%	100.0%
Methodology	100.0%	97.3%	100.0%	100.0%	100.0%	100.0%	100.0%
Findings	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
Conclusions	95.3%	100.0%	92.0%	93.8%	100.0%	97.8%	100.0%
Lessons Learned	97.7%	100.0%	100.0%	100.0%	97.4%	97.8%	98.2%
Emerging Good Practices	81.4%	89.2%	84.0%	84.4%	84.6%	84.8%	96.4%
Recommendations	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
Annexes	69.8%	59.5%	76.0%	65.6%	66.7%	58.7%	70.9%
Average	87.1%	88.5%	90.0%	89.5%	92.0%	91.3%	95.1%

This positive trend could be linked to the EVAL checklist for preparing the evaluation report, which was updated in 2021. All surveyed Evaluation Managers found the checklist to be "satisfactory" or "highly satisfactory" for its relevance, clarity, and applicability, and rated its helpfulness on average 80 on a scale from 0 (not helpful at all) to 100 (extremely helpful). This is an improvement on last year's survey results, in which helpfulness was rated on average 75 and some respondents found it "unsatisfactory". In this year's survey, 100% of the respondents noted that they frequently or always share this checklist with evaluators and experienced evaluators use the resource.

For the sample of 2020-2021 internal evaluations, an average of 86% of components were included, lower than independent evaluations and HLEs. 20% of internal evaluations did not include a 'title page using EVAL's template'. Only 40% of reports included the 'list of tables, figures, and charts' and only 50% provided all relevant 'Annexes'.

The 2020-2021 HLEs included on average 92.9% of components. 'Annexes' were incomplete for four out of six HLEs, while one out of six HLEs did not include a 'list of annexes' or 'criteria and questions'. However, contrary to independent evaluations, all HLEs included a 'list of tables, figures and charts'.

2.4. Summary High-Level Comparison of Independent, Internal and High-Level Evaluations

A direct comparison of the quality of independent, internal and high-level evaluations is not possible. This is because different QA tools were used to assess the quality. In addition, each of them required different levels of effort. However, the following provides general observations drawn from a high-level comparison.

Looking at overall ratings, both independent evaluations (2021) and high-level evaluations (2020-2021) have performed well with over 80% of reports reviewed receiving a "satisfactory" (5 out of 6) median rating. The sample of internal evaluations performed slightly lower with 70% of reports (or 7 out of 10) achieving a "satisfactory" median rating.

Similarly, none of the independent evaluations and high-level evaluations received a median rating lower than "somewhat satisfactory" (4 out of 6), while one internal evaluation received a median rating of "highly unsatisfactory" (1 out of 6) as it did not include the majority of necessary components.

As for specific components, a comparison of independent evaluations and internal evaluations shows that internal evaluations record more shortcomings for the components 'Project Background', 'Evaluation Methodology', and 'Findings'. A direct comparison with HLEs is not feasible as the components slightly differ from those of independent and internal evaluations.

2.5. Managing Evaluations during the COVID-19 Pandemic

Most evaluation reports considered for this report were designed and conducted during the COVID-19 pandemic. The aggregate QA results and survey data provided preliminary insights into how evaluation activities have fared in the context of the pandemic.

As noted in last year's QA Summary Report, several surveyed Evaluation Managers noted the impact of the global health crisis on the timeframe of evaluations. The average timeframe of independent evaluations conducted in 2021 is 0.37 months longer than the average timeframe of evaluations conducted in 2020. The median timeframe of evaluations conducted in 2021 increased to 4 months from 3 months in 2020. Despite the increased median and average timeframes, surveyed Evaluation Managers still rated the adequacy of the timeframe for completing all tasks associated with the evaluation on average 52 on a

scale from 0 (not at all adequate) to 100 (completely adequate), which is a drop from an average of 61 in last year's survey.

Evaluation Managers surveyed rated the adequacy of the quality of the evaluation process given the constraints of the pandemic on average 58 on a scale of 0 (not at all adequate) to 100 (completely adequate), which is the same average rating as in last year's survey. Of this year's respondents, 56% (9) considered the quality of the evaluation process as adequate while only 19% (3) considered it highly adequate. Evaluation Managers highlighted the methodological challenges that arose, and the difficulty of setting up virtual data collection missions to meet stakeholders and conduct participatory evaluation processes.

Nevertheless, a closer look at changes in satisfactory ratings (SS, S, HS) per item between 2020-2021, in comparison to 2019-2021, shows that most items did not just return to 2019 ratings; the 2021 ratings actually surpassed those of 2019. This suggests that the various measures taken by Evaluation Managers and EVAL, including updated and new tools and checklists and extending the timelines of evaluations, may have had positive effects on the overall quality of evaluation reports. In October 2020, EVAL released a "Protocol on collecting evaluative evidence on the ILO's COVID-19 response measures through project and programme evaluations". 33% (5) of them considered this protocol helpful and 47% (7) considered it extremely helpful to inform the task of managing evaluations, notably in the identification of the key evaluation questions.

The increased satisfaction and use of the guidelines may explain the significant improvement in the inclusion of evaluation questions to measure ILO's delivery in mitigation of the effect of the COVID-19 pandemic in independent evaluations. In collaboration with Universalia, EVAL added a specific item to the QA tool to assess the extent to which evaluation reports include evaluation questions that measure ILO's delivery in mitigating the effects of the COVID-19 pandemic (CE.9). The rating for this item improved from a "highly unsatisfactory" median rating (1) in 2020 to a "satisfactory" median rating (5) in 2021. Most internal evaluation reports reviewed were rated low on this item; however, half of the reports were conducted either before or very early on in the pandemic. There is no comparative data for high-level evaluations, since the HLE QA tool does not include this item.

2.6. SWAP

Universalia also analyzed the extent to which the ILO met UN-SWAP requirements, in 2021. The ILO is one of 69 organizations that are mandated to report against United Nations System-Wide Action Plan on Gender Equality and the Empowerment of Women (UN-SWAP-GEEW).

Reports are submitted on an annual basis using the UNEG endorsed Technical Note and related scorecard to report against the Evaluation Performance Indicator (EPI).³ In order to comply with this requirement, on an annual basis, ILO asks quality assurance (QA) consultants to rate the EPI contained in the scorecard in compliance with the instructions

³ UNEG. April 2018. UN-SWAP Evaluation Performance Indicator Technical Note.

found in the UN-SWAP-GEEW Technical Notes. The goal is for all UN system entities to “meet requirements” related to the EPI in terms of integrating GEEW in their respective evaluations.

By aggregating all ratings (0-3 scale) attributed to all 2021 evaluation reports on each of the three UN-SWAP criteria, the median score obtained across the 55 reports appraised corresponds to two (2), while the average score corresponds to 1.81, thus indicating that evaluation reports are, on average, satisfactorily conforming to UN-SWAP criteria.

The table below provides a detailed overview of the distribution of ratings across all reports, for each criterion. Assuming zero and one scores fall on the negative spectrum of the rating scale, while two and three ratings are positive, one can observe that most reports fall, on average, within the positive spectrum. Indeed, on average, 35% of reports fall on the negative spectrum (0 and 1), while the remaining 65% fall on the positive spectrum (2 and 3).

More specifically, most of the reports developed gender-responsive methodology, methods, tools and data analysis techniques (G2) (72%) and integrated GEEW in the evaluation scope (G1) (78%). However, the table also shows that 2021 evaluation reports fell short in integrating gender analysis in the findings, conclusions, and recommendations (G3). Fifty-five per cent of the appraised reports fall on the negative spectrum of the rating scale on that criterion.

Table 4.1 *Percentage of reports’ average ratings per UN-SWAP criterion for 2021*

UN-SWAP CRITERIA	NOT AT ALL INTEGRATED (0)	PARTIALLY INTEGRATED (1)	SATISFACTORILY INTEGRATED (2)	FULLY INTEGRATED (3)
G.1 GEEW is integrated in the evaluation scope of analysis and evaluation criteria and questions are designed in a way that ensures GEEW related data will be collected	5%	22%	56%	16%
G.2 A gender-responsive methodology, methods and tools, and data analysis techniques are selected	5%	16%	40%	38%
G.3 The evaluation findings, conclusions and recommendation reflect a gender analysis	2%	53%	40%	5%

► 3. Conclusions

Overall, there is a notable positive trend in ratings, across components, of independent evaluations since 2020. In 2021, all components of independent evaluations obtained “satisfactory” ratings in 2021 (median rating of 5 out of 6). This improvement is especially significant, as ratings saw a decrease for several components in 2020 largely due to the onset of the COVID-19 pandemic and its effects on the conduct of evaluations by ILO departments and offices. As noted in last year’s QA Summary Report, several surveyed Evaluation Managers noted the impact of the global health crisis on the timeframe of evaluations, the methodological challenges that arose, and the difficulty of setting up virtual data collection missions to meet stakeholders and conduct participatory evaluation processes.

A closer look at changes in satisfactory ratings (SS, S, HS) per item between 2020-2021 in comparison to 2019-2021 shows that most items did not just return to 2019 ratings but went above 2019 ratings in 2021. This suggests that the various measures taken by Evaluation Managers and EVAL, including updated and new tools and checklists and extending the timelines of evaluations, have likely had positive effects on the overall quality of evaluation reports. Nevertheless, Evaluation Managers continue to face constraints, as survey respondents raised concerns about navigating the complexities of the pandemic, including methodological and logistical limitations affecting the quality and reliability of data.

Despite the positive trend in overall ratings, there are notable shortcomings of independent evaluations that require continued attention across components. For instance, weaknesses in the executive summary of reports indicate a need for greater alignment with the EVAL Checklist 4.4: Preparing the Evaluation Report Summary. Reports still do not consistently integrate considerations for environmental sustainability and people with disabilities in the evaluation questions. Reports are also missing information on key methods, including types of analysis and sampling strategies, which could be further developed by integrating methodological elements mentioned in the inception report. Findings are not always sufficiently supported by evidence, especially the description of unintended outcomes. Overall, as in previous years, the linkages among evaluation questions, data collection methods, findings, conclusions, and recommendations could be made more explicit in many reports.

The review of a sample of 10 internal evaluations from 2020-2021 shows that the quality is more mixed across components, including project background, evaluation methodology and findings receiving median ratings lower than “satisfactory” (5). This indicates that the EVAL guidelines and checklists are less consistently used than for independent evaluations and HLEs.

As for 2020-2021 HLEs, there is also an overall improvement compared to the 2015-2019 sample reviewed in the previous year with median ratings for all components increasing or remaining the same, except for the evaluation background. While their overall quality is satisfactory for most components, shortcomings can be observed for the evaluation background and the evaluation methodology as well as specific items in other components, similar to independent evaluations.